REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 0704-0188

1. REPORT DATE *(DD-MM-YYYY)*	2. REPORT TYPE	3. DATES COVERED *(From - To)*
12/23/2013	Final Progress Report	01/01/10 - 09/30/13

4. TITLE AND SUBTITLE
Smart Distributed Sensor Fields

5a. CONTRACT NUMBER

5b. GRANT NUMBER
N00014-10-1-0477

5c. PROGRAM ELEMENT NUMBER

6. AUTHOR(S)
Saligrama, Venkatesh

5d. PROJECT NUMBER

5e. TASK NUMBER

5f. WORK UNIT NUMBER

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)
Trustees of Boston University
881 Commonwealth Avenue
Boston, MA 02215

8. PERFORMING ORGANIZATION REPORT NUMBER

9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)
Office of Naval Research
495 Summer Street
Suite 627
Boston, MA 02210-2109

10. SPONSOR/MONITOR'S ACRONYM(S)

11. SPONSOR/MONITOR'S REPORT NUMBER(S)

12. DISTRIBUTION/AVAILABILITY STATEMENT
Approved for Public Release; Distribution is Unlimited

13. SUPPLEMENTARY NOTES

14. ABSTRACT
Video cameras are critical to providing persistent surveillance capabilities for situational awareness. Currently, video analysis requires significant human supervision. Even many of the routine tasks ranging from detecting, identifying, localizing/tracking interesting events, discarding irrelevant data, to providing actionable intelligence currently requires significant human supervision. Human supervision is not scalable for providing persistent wide-area monitoring and particularly for monitoring a network of cameras that would be generally employed for theater- level operations. We develop methods for autonomous suspicious activity detection, multi-camera fusion and retrieval algorithms for large-scale WASdata

15. SUBJECT TERMS
Irregular and asymmetric warfare, WAS data, anomaly detection, search and retrieval, low bandwidth capability, low storage ability.

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	18. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON
a. REPORT	b. ABSTRACT	c. THIS PAGE			Venkatesh Saligrama
					19b. TELEPHONE NUMBER *(Include area code)*
U	U	U	UU	23	617-353-1040

Final Report

Project No: N00014-10-1-0477
Thrust: Asymmetric and Irregular Warfare

Smart Distributed Sensor Fields: Algorithms for Tactical Sensors

Submitted by

Principal Investigator: Venkatesh Saligrama, Professor
Department of Electrical and Computer Engineering
Boston University
8 St. Mary's Street, Boston, MA 02215
Phone: (617) 353-1040, Fax: (617) 353-6440
E-mail: srv@bu.edu

Submitted to

Technical Point of Contact: Dr. Martin Kruger, Program Officer
Office of Naval Research
ONR Department Code: 30
875 North Randolph Street
Arlington, VA 22203-1995

Duration: Jan 1, 2010 – Sept 30, 2013

Contents

1 Introduction

1.1 Scope

Video cameras are critical to providing persistent surveillance capabilities for situational awareness. Currently, video analysis requires significant human supervision. Even many of the routine tasks ranging from detecting, identifying, localizing/tracking interesting events, discarding irrelevant data, to providing actionable intelligence currently requires significant human supervision. Human supervision is not scalable for providing persistent wide-area monitoring and particularly for monitoring a network of cameras that would be generally employed for theater-level operations. The scope of this project is to develop new concepts for autonomous video analysis for highly cluttered urban environments.

1.1.1 Goals, Objectives & Challenges

We present techniques for autonomous and distributed operation of wide-area camera networks for video analysis in unstructured and highly cluttered urban environments. Our goals for video analysis are:

1. **Anomaly Detection:** Here we are interested in developing novel algorithms for autonomous real-time detection of suspicious activity.
2. **Search & Retrieval:** This task deals with pulling activities of interest in large wide-area surveillance video based on analyst generated input query.
3. **Multi-Camera Activity Fusion:** In this task we present algorithms for combining views from multiple cameras for dynamic scene characterization to improve detection, localization and tracking performance

The main challenges in developing concepts for video analysis is that urban scenarios provide a deluge of dynamic data. Identifying relevant information, such as meaningful change detection, in urban clutter is not easy. Second, doing so reliably, i.e., with small false alarms and missed detections is difficult and possibly impossible in harsh sensing environments (camera jitter etc). Third, combining views from multiple cameras for dynamic scene characterization and to improve detection, localization and tracking performance is hard. Finally, search and retrieval of activities of interest, indexing events of interest in long videos, is extremely challenging not only because of the inherent multi-scale nature of the phenomena.

1.2 Operational Naval Concept

Our main objective is to develop an essentially autonomous camera surveillance network in unstructured and highly cluttered urban & littoral environments. The new capabilities will include real-time abnormal activity detection, localization and tracking from multi-camera systems, and content summarization for fast indexing and search of archived video data. The operational performance improvements will include computationally efficient algorithms, demonstration in highly cluttered environments and improved ROC curves to improve reliability and robustness. The proposed effort deals with several of Navy's S& T focus areas in addition to Automated Image Analysis Asymmetric and Irregular Warfare, ISR, and Information Integration & Fusion.

2 Deliverables & Outcomes

The principle deliverable for this 6.1 project is this final report. Apart from this deliverable the project has resulted in several noticeable outcomes.

Transition Path The transition path for this project is towards building an *Agile Tactical SNET*. The search & retrieval concepts developed under this project has generated significant interest. To this end this project has led to a new 6.2 project for developing *Distributed Search Engines for Large Video Stores*. The search algorithm finds matches in a hash table and only retrieve video segments that meets search criteria. This is useful for remote video stores that generate significant data and there is insufficient bandwidth to transmit this data in a timely manner. We have partnered with **NRL-Stennis** to help transition our concepts onto their CisView platform. Our transition plan in the new 6.2 project includes code development with updates targeted for delivery to ONR each year.

Papers & Reports This project has produced a number of publications in internationally recognized conferences and reputed journals. We have also presented this material in a number of universities and parts of this work has appeared as chapters in edited books. We list these publications here:

1. G. Castanon, P.-M Jodoin, V. Saligrama, A. Caron *Activity Retrieval in Large Surveillance Videos*, Elsevier E-reference for Signal Processing, 2013
2. V. Saligrama, Z. Chen, *Video Anomaly Detection Based on Local Statistical Aggregates*, IEEE Computer Vision and Pattern Recognition (CVPR), 2012
3. G. Castanon, V. Saligrama, P. M. Jodoin, A. Caron, *Exploratory Search in Long Surveillance Videos*, ACM Multimedia, 2012 (full paper, acceptance rate: 20
4. Y. Benezeth, P. Jodoin, V. Saligrama, *Abnormality Detection Using Low-Level Co-occurring Events*, Pattern Recognition Letters, 2011
5. P.M. Jodoin, V. Saligrama, J. Konrad, *Behavior Subtraction: A new tool for Video Analytics*, IEEE Transactions on Image Processing, Sept 2012
6. V. Saligrama, J. Konrad, P. M. Jodoin, *Video Anomaly Identification*, IEEE Signal Proc. Magazine, 2011
7. E. Ermis, P. Clarot, P. M. Jodoin, V. Saligrama, *Activity Based Matching in Distributed Camera Networks*, IEEE Transactions on Image Processing, Sept 2010
8. E. Ermis, V. Saligrama, P. Jodoin, *Information Fusion and Anomaly Detection with Uncalibrated Cameras in Surveillance*, in Multimedia Information Extraction, M. Maybury (eds), IEEE Press, 2012
9. Y. Benezeth, P. Jodoin, V. Saligrama, *Modeling Patterns of Activity and Detecting Abnormal Events with Low Level Co-occurrence of activity*, in Distributed Video Sensor Networks, Bir Bhanu et. al. (eds), Springer 2011

3 Detailed Technical Approach & Results

Video surveillance has been an area of significant interest in both academia and industry. We develop a novel event-based framework for anomaly detection, multi-camera fusion and activity

retrieval. Our approach is based on statistical learning techniques for video analysis. At a fundamental level this requires three steps, namely,

1. Feature Selection & Extraction: The main goal here is to select descriptors that are not only informative but also have sufficiently low complexity such that they are robust, relatively easy to extract, and amenable to real-time analysis. For instance, tracks are high dimensional features that are difficult to extract in cluttered scenarios. Our goal is to select informative low-dimensional features that are robust to photometric properties and relative easy to extract.
2. Feature Modeling: The goal here is to develop probabilistic models to characterize dynamic evolution of features over space and time.
3. Video analysis: This involves algorithms for anomaly detection, multi-camera fusion and retrieval.

3.1 Anomaly Detection

Anomaly detection for video surveillance has gained importance [2, 3, 6, 11, 15, 25, 29, 32, 33, 35, 37, 57]. Our focus is on problems, where we are given a set of nominal training videos samples. Based on these samples we need to determine whether or not a test video contains an anomaly. We consider anomalies in motion attributes. Such outliers can include (un)usual motion patterns of (un)usual objects in (un)usual locations. These encompass anomalies such as dropped baggage, illegal U-turns, and sudden movements.

We focus on anomalies that have local spatio-temporal signatures. The work reported here has appeared in our CVPR 2012 paper [43]. By locality we mean that the spatio-temporal region surrounding the anomalous region appears to follow the nominal activity and carries little information about the anomaly itself. For instance, the appearance of a bicyclist as shown in Fig. 1 illustrates spatio-temporal locality. As is seen outside a small window in time or in space the optical flow magnitudes look remarkably similar to nominal activity. We also consider other cases where locality is only temporal. These include cases such as sudden crowd movement [1] or illegal U-turns [6] We exploit these ideas by proposing a statistical non-parametric notion of locality and derive data-driven rules for anomaly detection with predictable performance and statistical guarantees. Our approach is related to a number of other non-parametric data-driven approaches such as [46, 63] with important differences. Existing statistical approaches do not account for local anomalies, i.e., anomalies that are localized to a small time interval and/or spatial region. Our statistical locality notion leads to an elegant characterization of anomaly detection and suggests novel empirical rules. A fundamental insight gained from our theoretical results is that the optimal decision rules for local anomalies are local irrespective of the global statistical dependencies exhibited in the nominal behavior. This key insight implies that the inherently large ambient data dimension is inconsequential. Our local empirical rules fuse local statistics and produce a composite score for a video segment. Anomalies are declared by ranking composite scores for video segments. Our anomaly detection algorithm is described in Fig. 2. Our setup extracts local low-level motion descriptors and resembles other common approaches. Adam et al. [2] use histograms of optical flows at specific "local monitors" to derive decision rules for anomaly detection at those locations. Itti and Baldi consider low-level feature descriptors at every location [27] and use possion statistics for modeling nominal activity.

We propose a joint probability distribution of the low-level motion descriptors under nom-

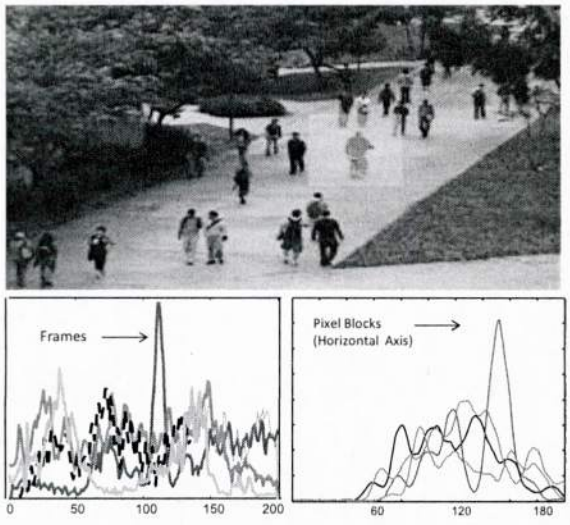

Figure 1: Illustration of local anomaly. Top: Illustrates frame of a video segment [35] with anomaly (bicycle). Bottom Panel (Left): Optical flow magnitude averaged over the red block vs. frame number for nominal and anomalous video segments. (Right): Optical flow magnitude averaged over different blocks along horizontal pixel blocks for different nominal and anomalous video segments.

inal as well as anomalous distributions. Such joint distributions have also been considered extensively. Kim et al. [32] also extract local optical flow and enforce consistency across locations through Markov Random Field models. Benezeth et al. [6] use binary background subtraction to extract motion labels and then model these local features using a 3D Markov Random Field (MRF). Kratz et al. [33] extract spatio-temporal gradient to fit Gaussian model, and then use HMM to detect abnormal events. Mahadevan et al. [35] model the normal crowd behavior by mixtures of dynamic textures.

We introduce novel structural assumptions on the joint distributions to account for spatial and temporal locality of anomalies. Our locality assumption leads us to consider statistics on local 3D brick patches (space-time blocks) across different overlapping locations. These statistics are obtained through spatio-temporal filters as shown in Fig. 2. Our 3D modeling superficially resembles Boiman and Irani [11] but is different. They consider ensembles of 3D bricks and derive Gaussian models for matching test ensembles at a specific location with corresponding ensembles in a database. However, our goal is statistical and does not attempt to match 3D bricks at a location. Rather (see Fig. 7) we first compute location specific K-nearest neighbor (NN) distance for each 3D brick. We then normalize and compute a composite score by aggregating weighted K-NN distances from all the locations. This composite score is ranked against other such composite scores associated with training video segments. We then declare low scores as anomalies. It turns out that fusing local 3D brick statistics in this manner has theoretical significance. The empirical composite scoring and ranking scheme asymptotically converges to the optimal decision rule for maximizing detection power subject to false alarm constraints.

Our work is also related to Cong et. al. [15] who consider dictionary learning methods. There 3D patches with specific temporal and spatial scale are chosen to match each scenario. A dictionary of representative patterns are learnt based on training video. Anomalies are declared

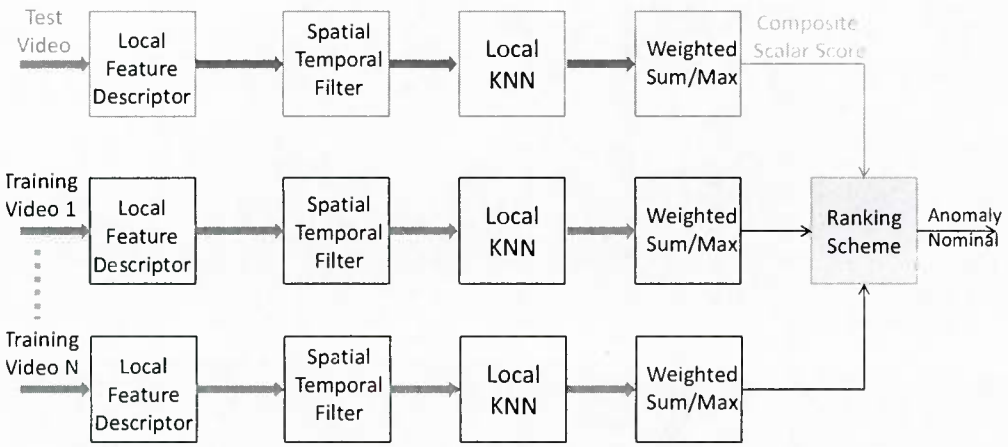

Figure 2: Overview of Anomaly Detection Algorithm. Motion descriptors are first extracted and quantized into small blocks. Spatio-Temporal filters at different scales are applied to obtain smooth estimates at each spatio-temporal location for each feature descriptor. Local KNN distance for each location is computed for training and test video. These local KNN distances are aggregated to produce a composite score for the test and training video. The composite scores are ranked to determine anomalies.

if the test sample cannot be represented using a sparse set of dictionary patterns. It is worth mentioning that we could incorporate their ideas into our scheme. Sparse decomposition for each spatio-temporal scale can be viewed as a feature vector that feeds into our local KNN block (see Fig. 7).

Other work on video anomaly detection includes social force models by Mehran et. al. [37]; Normalized cut clustering by Zhong et. al. [64]; and trajectory based methods [3, 25, 57].

3.1.1 Video Locality Model and Feature Descriptors

A video snippet x is typically a short segment of video. Training data can consist of several snippets, $x^{(1)}, x^{(2)}, \ldots, x^{(n)}$. For theoretical purposes we assume that the different snippets are independent of each other. These snippets can be obtained by partitioning a longer video into short non-overlapping segments.

For a video snippet, x, we associate a graph $G = (V \times T, E)$. The set V is associated with spatial locations and the set T is associated with temporal locations in the video snippet. Each location, $v \in V$ and time $t \in T$ is associated with a feature descriptor $x_{v,t}$. While it is theoretically possible to consider all pixel locations and temporal instants, we quantize into $10 \times 10 \times 5$ non-overlapping blocks. We call these blocks as atoms and we associate average values of features for each atom. Two atoms are connected if they are either temporal or spatial neighbors. The rest of development with regards to Mask and Markov assumptions follow as in the previous section (also see Fig. 4).

Feature Descriptors: We now describe local features that are associated with each node (atom) of our graph. During feature extraction we compute a feature value for each pixel. Then, the pixel-level features are condensed into a multi-dimensional vector for each atom by averaging each feature component over all the pixels within the atom. We use the following local features:

(1) Persistence: Activity is detected using a basic background subtraction method (as for instance in [6]). The initial background is estimated using median of several hundred frames. Then, the background is updated using the running average method. We flag each pixel as part of the background or foreground. Persistence, for an atom, is the percentage of foreground pixels in the atom.
(2) Direction: Motion vectors are extracted using Horn and Schunck's optical flow method [9]. Motion is quantized into 8 directions and an extra "idle" bin is used for flow vectors with low magnitude. The feature for each atom is a 9-bin *un-normalized* motion histogram. The value for each bin corresponds to the number of pixels moving in the direction associated with the bin.
(3) Motion Magnitude: Magnitude of motion vectors for each bin (except the idle bin) is computed and averaged over all the pixels in the atom.

We thus have an 11-dimensional descriptor for each atom. While our setup is sufficiently general and admits other descriptors we use only these components.

3.1.2 Algorithm for Video Anomaly Detection

Recall we are given training video samples and a test video sample. To reduce real-time delay we breakup the test video sample into test video snippets, $\eta^{(1)}, \eta^{(2)}, \ldots, \eta^{(m)}$. Our task is to determine which of the test snippets contain an anomaly. For convenience, we partition training video in to snippets $(x^{(1)}, \cdots, x^{(n)})$ each of the same length as a test snippet η. Our algorithm consist of three steps:
(1) Local Scores: For any snippet y, which denotes either a test or training snippet, a local score at spatial location v, temporal instant, t, and at spatio-temporal scale s, is computed (see Algorithm 1). We choose an averaging spatio-temporal filter for simplicity in Algorithm 1.

Algorithm 1 Score for y at location (v, t), at scale s.

Input: Training Descriptors, $\{x_{u,\tau}^{(j)}\}$, $\forall\, j$, u, τ; K for KNN
Output: $d_{y,t}(s)$

1: Filter at scale s: $x_{v,\tau}^{(j)} \leftarrow Filter_s(x_{u,\tau}^{(j)})$
2: Distance Computation: $d_{v,t,\tau,j} = d(y_{v,t}, x_{v,\tau}^{(j)})$, $\forall\, j$
3: Compute $d_{v,t,(\ell)}$ the ℓth nearest neighbor distance by sorting $d_{v,t,\tau,j}$.
4: Average:$d_{v,t} \leftarrow \frac{1}{K} \sum_{\ell=K+1}^{2K} d_{v,t,(\ell)}$
5: Normalize $d_{y,t} \leftarrow \frac{d_{v,t}}{D_v}$; where $D_v = \max_t d_{v,t}$

(2) Snippet Score: Compute composite score for snippet, y, from local scores obtained in Algorithm 1:

$$d_y(s) = \max_{v,t} \bar{d}_{v,t}(s)$$

(3) Anomaly Detection: Rank test snippet, η at scale s:

$$R_s(\eta) = \frac{1}{n} \sum_{j=1}^{n} \mathbb{I}_{\{d_{x^{(j)}}(s) > d_\eta(s)\}}$$

Note that our feature descriptors–magnitude, direction, persistency—have different dynamic ranges. Here we ranked separately with respect to the different descriptors. Anomalies are

declared if the rank for any descriptor falls below the desired false alarm threshold. Anomaly is localized by identifying the spatial and temporal locations in the snippet that contribute towards achieving the rank, $R_s(\eta)$.

Tuning Parameters: Our algorithm requires only two parameters, namely, K for KNN distance computation and scale s. It turns out that our results are generally robust to a wide range of K and is not an issue. In all our simulations we choose K to be about 50. Scale s can be dealt with in two possible ways: (**1**) Compute ranks over different scales and declare anomaly if the rank at some scale falls below the threshold. This procedure is conservative; Nevertheless, it controls false alarms at desired level asymptotically. (**2**) Use context to determine sensible temporal and spatial scales. This idea has been used before by Cong et. al. [15], who choose appropriate basis depending on the scenario. We choose small scales if small scale anomalies (abandoned or unusual objects) are important and choose larger scales for spatial anomalies such as U-turns or global change in behavior.

Computational Issues: KNN distance computation is our main bottleneck. It scales linearly with the number of 3D bricks. To overcome this drawback recent approaches for computing approximate nearest neighbors based on locality sensitive hashing(LSH) [4] can be used. While we do not present results based on LSH here, in our preliminary experiments we have noticed that it can drastically reduce the computation time (scaling as fourth root of the number of 3D bricks) with little loss in performance.

3.1.3 Experimental Results

The UCSD Ped1 dataset [54] contains 34 training clips of nominal patterns and 36 testing clips of various abnormal events, for example, bicycles, skaters, carts, etc. Each clip has 200 frames (20 seconds), with a 158×238 resolution. The challenge in this dataset is that the scenes are extremely crowded. To apply our algorithm, first we calculated optical flow and aggregated optical flow into histogram and magnitude features. We divided the videos into overlapping spatio-temporal blocks of 30pixels×20pixels×5frames (the block size was chosen such that each block does not contain too many objects which may interfere with one another) and then we applied our algorithm on snippets consisting of 5 frames. We also experimented with larger snippets and noticed little performance degradation.

Some image results are shown in Figure 3. Our algorithm can detect different types of anomalies. We compared our method with SRC proposed in [15] and MDT proposed in [35]. We also compared our method with Social force and MPPCA, etc. We found that [43] our method outperforms all the other algorithms. In Table 1, some evaluation results are presented: the Equal Error Rate (EER) (ours 16% < 19% [15]), and Area Under Curve (AUC) (ours 92.7% > 86% [15]). From these comparisons, we can conclude that our algorithm outperforms other state-of-the-art algorithms. One additional advantage of our algorithm is that while providing frame level results, we can also provide anomaly localization by back-tracing to the block with max statistics.

3.2 Multi-Camera Processing

We have developed algorithms for finding pixel level correspondences between multiple cameras that have partially overlapping field of views. Our problem is motivated by the wide area surveillance applications. We present algorithms for settings where cameras have significantly

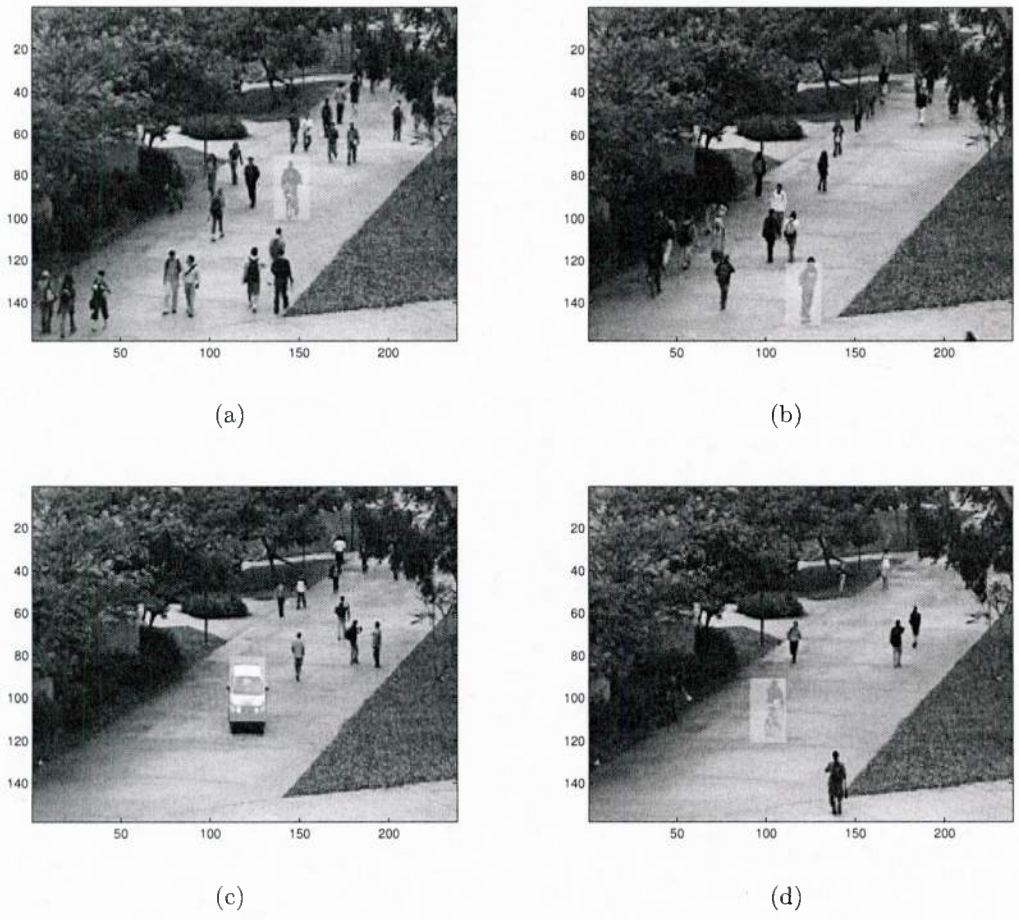

Figure 3: Abnormal event detections for UCSD Ped1 datasets. The objects such as cars, bicycles, skaters are all well detected.

different orientations and zoom levels with respect to the scene. We propose a correspondence method based on activity features that, unlike photometric features, have certain geometry independence properties. The proposed method works is directed towards general surveillance scenarios, where prior calibration is not possible. In addition we sced techniques which require little processing power, and is communication resource aware.

Tracking moving objects in video is a difficult problem and has been approached from many different angles. Some of the most successful techniques are particle filtering [26] and covariance-matrix techniques [40,53]. The latter techniques have proved robust to size scaling, pose change, illumination variations, while at the same time can be efficiently implemented using the integral image concept. The main challenges in their adoption will be in identifying covariance features to use (e.g., luminance and color can be tracked jointly with position, structure, velocity, etc.) and extending the approach to multiple cameras. While classical multiple-camera object tracking, dominant in computer vision literature, hinges on frame-to-frame correspondence between cameras, we propose to establish such correspondence using only the event-based representations [19, 20]. Dynamic events, unless occluded, leave a unique signature in camera-

Method	EER	AUC
MPPCA [35]	40%	59%
SF [35]	31%	67.5%
MDT [35]	25%	81.8%
Sparse [15]	19%	86%
Ours	**16%**	**92.7%**

Table 1: Quantitative comparison of our algorithm with [15] and [35]. EER is equal error rate and AUC is the area under ROC.

acquired views regardless of the projection angle. The figure below demonstrates a number of correspondence experiments conducted. The task is to determine the mapping between locations across different cameras that share an overlapping field of view.

While a car traveling on a highway induces similar luminance/color pattern changes on cameras viewing it from different angles, it also induces similar patterns of dynamic events (e.g., sequences of idle and busy periods) across views. By seeking to associate dynamic events, instead of brightness patterns, across different views we bypass the difficult issues related to 3-D geometry of viewing angles and high-bandwidth requirements. We expect the event-based correspondence to be also helpful in activity recognition on account of multiple sources observing the same target. A detailed description of this approach has appeared in [21]

The general method heretofore is the so called scale invariant feature transform (SIFT) based method. The main difficulty is that when the cameras have significantly different orientations, SIFT method fails to produce meaningful results. However, activity features are geometrically independent demonstrating the effectiveness of the proposed method for a large class of surveillance scenarios.

3.3 Search & Retrieval

The problem of exploratory search is motivated by the need for searching large video stores that are produced remotely. We are interested in finding activities that match a wide variety of queries. The need for this technology at this time is to enable Asymmetric and Irregular Warfare capability. Specifically, this aspect of the project is aimed at enabling analytics for wide area imagery. The goal is to enable searching of suspicious activity in cluttered environments while using low bandwidth and low storage in conducting search. This capability will allow a remote user to search terabytes of wide area imagery efficiently by operating on a compressed data space such as a hash table and enable pulling video that meets search criteria.

3.3.1 Challenges

The main challenges that arise in an Exploratory Search system in large-scale surveillance videos are listed below:

1.) Data lifetime: since video is constantly streamed, there is a perpetual renewal of video data. This calls for a model that can be updated incrementally as video data is made available. The model must also scale well with the temporal mass of the video.

2.) Unpredictable queries: the nature of queries depends on the field of view of the camera, the scene itself and the type of events being observed. The system should support queries of

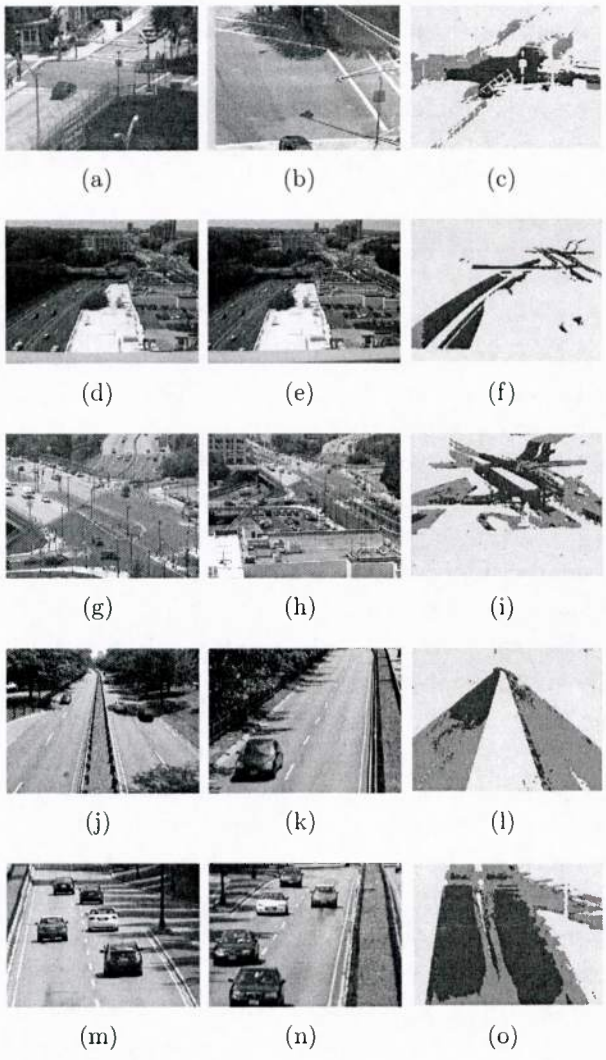

Figure 4: Occlusion map using left-right check and the proposed method: (a,d,g,j,m) Camera 1 frames, (b,e,h,k,n) Camera 2 frames, (c,f,i,l,o) Segmentation results for Camera 1 frames: Red regions appear in Camera 1 frame but not in Camera 2 frame, blue regions are common in Camera 1 and Camera 2, green regions carry no motion.

different nature that can retrieve both recurrent events such as people entering buildings and infrequent events such as cars performing U-turns, cars passing, car dismounts etc.
3.) Unpredictable event duration: events are unstructured. They start anytime, vary in length, and overlap with other events. The system is nonetheless expected to return complete events regardless of their duration and whether or not other events occur simultaneously.
4.) Clutter and occlusions: Tracking and tagging objects in urban videos is challenging due to occlusions and clutter; especially when real-time performance is required.
5.) Airborne Videos: Airborne videos offer new challenges. They record wide-area imagery and use mega-pixel cameras. On the other hand due to their constant motion registration is required before the footage can be subjected to video analysis. Imperfect registration is a common issue and this requires methods that are robust to imperfections introduced in the registration process.

Related Work:
Image Based Approaches: There has been significant work in the literature on indexing images. Nevertheless, these approaches run into immediate problems in video search: the size of the data representation is of the same order (or larger) than the images themselves. Furthermore, some traditional approaches are based on image retrieval, which relies on matching a bag of features. In contrast video activities have a time component. Therefore, a meaningful match must not only be matched at the frame level but also coherent in time to capture semantically meaningful spatio-temporal relationships. Most video papers devoted to summarization, search and retrieval focus on broadcast videos such as music clips, sports games, movies, etc. These methods typically divide the video into "shots" [18, 48, 50, 51] by locating and annotating key frames corresponding to scene transitions. The search procedure exploits the key frames content and matches either low-level descriptors [50] or higher-level semantic meta-tags to a given query [65]. Unfortunately, surveillance videos are fundamentally different from conventional videos. Most surveillance videos often contains many unrelated activities and events and so surveillance videos cannot be decomposed into "scenes" separated by key frames that one could summarize with some meta tags or a global mathematical model. Furthermore, surveillance video have no closed-caption or audio track one could rely on [65].
Video Clustering Based Approaches: This motivates approaches that attempts to index the dynamic content of the video in a way that is compatible with arbitrary upcoming user-defined queries. In that perspective, most scene-understanding video analytic methods work on a two-stage procedure: (1) learn patterns of activities via some clustering/learning procedure and than (2) recognize new patterns of activity via some classification stage. Since activities in public areas often follow some basic rules (think of traffic lights, highways, building entries, etc) the training stage often quantifies space and time into a number of states with transition probabilities. Common models are HMMs [34, 38, 41, 55, 62], Bayesian networks [12, 59], context free grammars [56], and other graphical model [36, 49, 57]. As for the classification stage, it is either used to recognize pre-defined patterns of activity [8, 10, 16, 23, 28, 47, 49, 60, 61] (useful for counting [16, 52]) or detect anomalies by flagging everything that deviates from what has been previously learned [5, 14, 25, 39, 41, 42, 45]. We also note that methods working on global behavior understanding often rely on tracking [12, 36, 41, 55, 58] while those devoted to isolated action recognition relies more on low-level features [10, 17, 23, 47, 61]. Although these methods could probably be tuned to index the video and facilitate search, very few papers explicitly address this question. One such paper is the one by Wang *et al.* [57]. There method decomposes

the video into *clips* in which the local motion is quantized into *words*. These words are than clustered into so-called *topics* such that each clip is modeled as a distribution over these topics. Queries being a combination of these topics, their search algorithm fetches every clip containing all of the topics mentioned in the query. A similar approach can be found in [24, 34, 59]

But search techniques focused on global explanations operate at a competitive disadvantage: the preponderance of clutter (requirement four) in surveillance video makes the training step of scene understanding prohibitively difficult. Second, since these techniques often focus on understanding recurrent activities, they are unsuitable for retrieving infrequent events - this can be a problem, given that queries are unpredictable (requirement two). Finally, the training step in scene understanding can be prohibitively expensive, violating requirement three, large data lifetimes.

3.3.2 Search Algorithm

The concepts developed here accounts for the challenges posed by search in large surveillance videos and overcomes some of the drawbacks of traditional approaches. First, we extract a full set of features as we have no a priori knowledge of what query will be asked. Unlike scene understanding techniques, we have no training step; this would be incompatible with the data lifetimes and magnitudes of the corpus. Instead, we develop an approach based on exploiting temporal orders on simple features, which allows us to find arbitrary queries quickly while maintaining low false alarm rates. The results reported here is based on work that has already resulted in a number of publications [7, 13, 19, 20, 30, 31, 44].

We tackle the aforementioned challenges through the following sub-components:

Efficient Representation: First, a capability for efficient representation and storage of the data must be developed; preferably, one which is smaller than the video, as even simple surveillance video can represent terabytes of data. The stored video must be sufficiently informative so that activities that could potentially of interest is preserved in the compressed representation. In order to address these challenges arising from efficient representation, we propose to employ simple pixel-level features (Motion, Size, Color, Persistence), and rely on their spatio-temporal relationships to identify query matches. We propose to divide up incoming video into space-time cubes (and pyramids of those cubes, called trees), and compute amalgamations of pixel-level features for each cube and pyramid as illustrated in Fig 2. Each of these feature trees is hashed using locality sensitive hashing (LSH) for fast retrieval. This approach yields a highly efficient representation of the video; a 5-hour, 7 GB video is compressed to a 5 MB index.

Query Representation: A query pattern must be generated, either directly or via exemplars, to search through the video. This query pattern must be sufficiently descriptive to account for semantically meaningful spatio-temporal relationships. An example of such a query interface is illustrated in Fig. 6.

Search Algorithm: The search algorithm must efficiently find semantically meaningful matches to its query pattern in its data representation. To describe the search engine we will first present a block diagram view of the overall system. The main idea is to reduce the problem to the relevant data, and then reason intelligently over that data. This process is shown in Fig. 7. As data streams in, video is pre-processed to extract relevant features - activity, object size, color, persistence and motion.These low-level features are hashed into a fuzzy, light-weight lookup table by means of LSH [22]. We propose LSH because it can account for spatial variability and

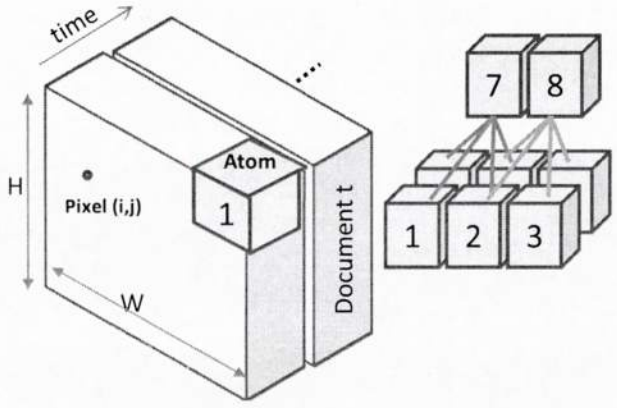

Figure 5: (Left) Given an $W \times H \times F$ video, *documents* are non-overlapping video clips each containing A frames. Each of the frames are divided into *tiles* of size $B \times B$. Tiles form an *atom* when aggregated together over A frames. (Right) Atoms are grouped into two-level trees - every adjacent set of four atoms is aggregated into a parent, forming a set of partially overlapping trees.

reduces the search space for a user's query to the set of relevant partial (local) matches.

Our search engine optimizes over the partial matches to produce *full matches*; segments of video which fit the entire query pattern, as opposed to part of it. This optimization operates from the advantageous standpoint of having only to reason over the partial matches, which are the relevant subset of the video. In surveillance video, where a long time can pass without relevant action, this dramatically reduces the workload of the optimization algorithm. The first is a greedy approach which flattens the query in time. Second, a novel dynamic programming (DP) approach, exploits the causal ordering of component actions that makeup a query. DP reasons over the set of partial matches and finds the best full match.

In our work we have developed GUIs to describe queries. The user enters the number of action components which the analyst wishes to find, and then draws the motion patterns for those actions. In order to recognize the complete set of actions in the video, we first get the set of matches to each individual action component. Because we have the video components hashed, this is an incredibly quick lookup - it is linear in the number of matching components. This is a convenient property, because action is frequently sparse in a video, and so scaling with the number of matches makes the actual length of the video irrelevant for performance. All that matters is the amount of action in the video. Once we are given a set of matches for each action component, the search for a full match can be formulated as a dynamic programming problem. We employ the Smith-Waterman algorithm for genome-matching to find the ranked set of matches in the video.

Our system currently supports a combination of motion and object type queries. In the follow on project we would like to extend our approach to support longer term activities where the routes or motion attributes could be uncertain. These type of queries could involve activities corresponding multiple destinations over large time scales. We are also currently investigating how to extend our GUI based querying to represent complex motion patterns. In this context we are planning on moving out of the realm of manually-created queries into exemplar-based querying. Sometimes, the identifying structure of an action may be difficult to decipher for a user, but they could provide a number of examples ("I know it when I see it"). We propose to

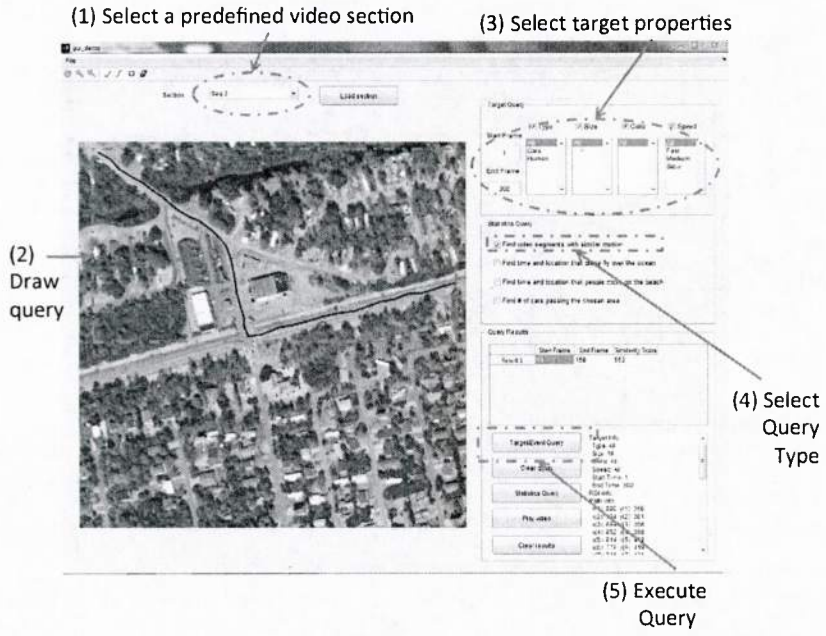

Figure 6: The query creation GUI provides a straightforward way to construct queries. The user draws each action component (shown in blue), and can additionally specify features.

Task	Frames	Resolution	# queries	Retrievals	False Retrievals	Groundtruth	Time/query
1	500	3850×5950	29	33	1	33	1.73 sec.
2	500	1951×5950	11	11	0	11	0.052 sec.

Table 2: Results for processing the Airborne data. Car queries representing routes as long as 1500 pels were examined. The specs of the computer used is Intel Core i5, @ 2.67 GHz 2.66 GHz, 4.0GB RAM.

break down exemplar videos to produce action components for search.

3.3.3 Results

Table 2 summarizes the results obtained from processing the Airborne data. We examined routes as longs as 1500 pels. Examples of somes of the examined routes are shown in Fig. 8. Some of those routes undergo strong occlusion and others have many turns. Fig. 9 shows the ROC for the tasks 1 and 2. The Retrieval Score Threshold is the minimum path score (generated by Algorithm) required in order to declare this path as a search result. The points on Fig. 9 represent 20 different Retrieval Score Threshold values in the range of 0:1:20 (MATLAB notations). As seen by the generated ROC, our technique performs well by generating 0.85 correct detection rate with 0.1 false alarms.

This approach represents a fundamentally different way of approaching the video search problem. Rather than relying on an abundance of training data or finely-tuned features to

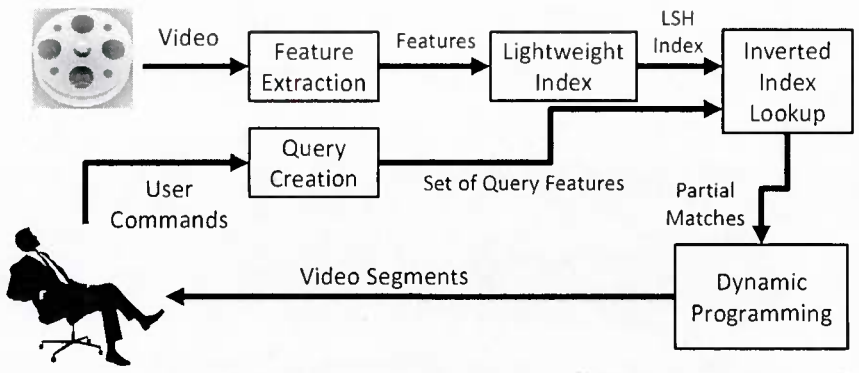

Figure 7: From streaming video, low level features for each document are computed and inserted into a fuzzy, lightweight index. A user inputs a query, and partial matches (features which are close to parts of the query) are inserted into a dynamic programming (DP) algorithm. The algorithm extracts the set of video segments which best matches the query.

differentiate actions of interest from noise, we rely on simple features and causality. In addition to the clear benefits in terms of a run-time which scales sub-linearly with the length of the video corpus, the simple features and hashing approach render the approach robust to user error as well as poor-quality video. The results demonstrate clearly that causality and temporal structure can be powerful tools to reduce false alarms. Another added benefit is how the algorithm scales with query complexity. Whereas algorithms such as topic modeling or a feature-based matching suffer as queries becomes more complex due to efforts to characterize the query, the two-step approach becomes more successful - the more action components in a query, the more likely it is to differentiate itself from noise. There is, of course, non-temporal structure that we have yet to exploit. Spatial positioning of queries, such as "The second action component must occur to the northeast of the first", or "The second action component must be near the first" is a simple attribute which may further differentiate queries of interest from background noise. This is not to say that the approach is not without its limitations. It requires that the activity being described contain discrete states, each of which is describable by a simple feature vocabulary. Complex actions like sign language or actions which are to fast or too small to be identified at the atom level will be difficult to search for.

References

[1] Unusual crowd activity dataset. `http://mha.cs.umn.edu/Movies/Crowd-Activity-All.avi`.

[2] A. Adam, E. Rivlin, I. Shimshoni, and D. Reinitz. Robust real-time unusual event detection using multiple fixed-location monitors. *IEEE Trans. Pattern Anal. Machine Intell.*, 30(3):555–560, 2008.

[3] S. Ali and M. Shah. Floor fields for tracking in high density crowd scenes. *ECCV*, 2008.

[4] Alexandr Andoni and Piotr Indyk. Near-optimal hashing algorithms for approximate nearest neighbor in high dimensions. *Commun. ACM*, 51:117–122, January 2008.

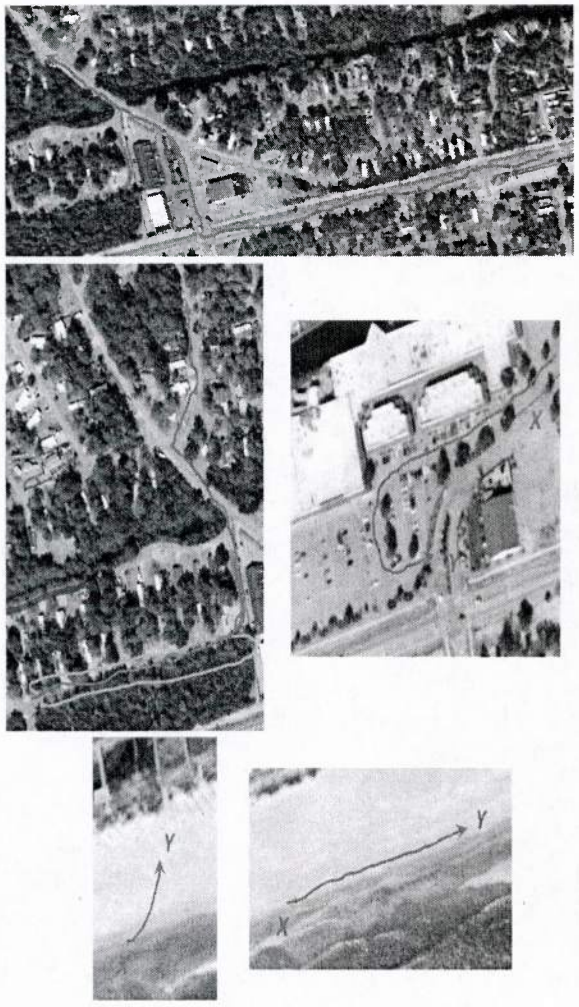

Figure 8: Examples of the examined routes. Routes are shown in red and they start from point X and end at point Y. Some of the routes undergo strong occlusion (see blue region, second row, left) and others undergo many turns (see second row, right).

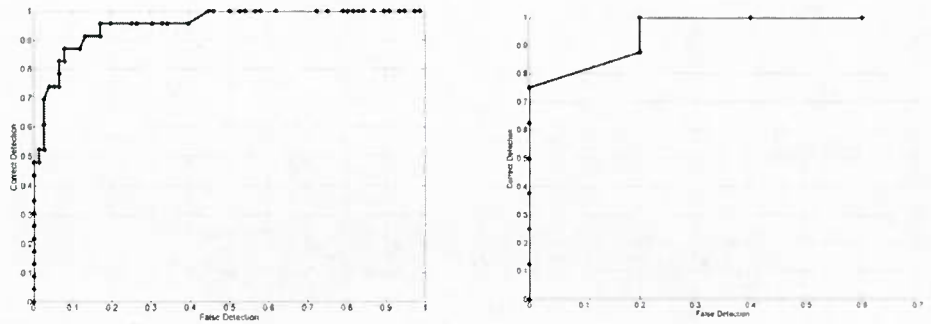

Figure 9: ROC for the examined airborne data. Here results for task 1 and 2 are on the left and right respectively. The points on the graph represent different values for Retreival Score Threshold.

[5] A Basharat, A. Gritai, and M. Shah. Learning object motion patterns for anomaly detection and improved object detection. In *Proc. IEEE Conf. Computer Vision Pattern Recognition*, pages 1–8, 2008.

[6] Y. Benezeth, P. Jodoin, V. Saligrama, and C. Rosenberger. Abnormal events detection based on spatio-temporal co-occurences. *CVPR*, 2009.

[7] Y. Benezeth, P. M. Jodoin, V. Saligrama, and C. Rosenberger. Abnormal events detection based on spatio-temporal co-occurences. In *Proc. IEEE Conf. Computer Vision Pattern Recognition*, pages 2458–2465, 2009.

[8] M. Bennewitz, G. Cielniak, and W. Burgard. Utilizing learned motion patterns to robustly track persons. In *Proc. IEEE Int. Workshop on Vis. Surv. and Perf. Eval. of Tracking and Surv.*, pages 1–8, 2003.

[9] B.Horn and B. Schunck. Determining optical flow. *Artif. Intell.*, 17(1-3):185–203, 1981.

[10] A. F. Bobick and J. W. Davis. The recognition of human movement using temporal templates. *IEEE Trans. Pattern Anal. Machine Intell.*, 23(3):257–267, 2001.

[11] O. Boiman and M. Irani. Detecting irregularities in images and in video. *ICCV*, 2005.

[12] S. Calderara, R. Cucchiara, and A. Prati. A distributed outdoor video surveillance system for detection of abnormal people trajectories. In *Proc. IEEE Conf. on Dist. Smart Cameras*, pages 364–371, 2007.

[13] G. Castanon, V. Saligrama, P.M. Jodoin, and A. Caron. Exploratory search in long surveillance videos. In *ACM Multimedia Conference*, 2012.

[14] C. Chuang, J-W Hsieh, and K-C Fan. Suspicious object detection and robbery event analysis. In *ICCCN*, pages 1189–1192, 2007.

[15] Y. Cong, J. Yuan, and J. Liu. Sparse reconstruction cost for abnormal event detection. *CVPR*, 2011.

[16] R. Cutler and L. S. Davis. Robust real-time periodic motion detection, analysis, and applications. *IEEE Trans. Pattern Anal. Machine Intell.*, 22(8):781–796, 2000.

[17] Q. Dong, Y. Wu, and Z. Hu. Pointwise motion image (PMI): A novel motion representation and its applications to abnormality detection and behavior recognition. *IEEE Trans. Circuits Syst. Video Technol.*, 19(3):407–416, 2009.

[18] A. Doulamis and N. Doulamis. Optimal content-based video decomposition for interactive video navigation. *IEEE Trans. Circuits Syst. Video Technol.*, 14(6):757–775, 2004.

[19] E.B. Ermis, V. Saligrama, P.-M. Jodoin, and J.Konrad. Abnormal behavior detection and behavior matching for networked cameras. In *ACM/IEEE Int. Conf. Distributed Smart Cameras*, September 2008.

[20] E.B. Ermis, V. Saligrama, P.-M. Jodoin, and J.Konrad. Motion segmentation and abnormal behavior detection via behavior clustering. In *Proc. IEEE Int. Conf. Image Processing*, October 2008.

[21] Erhan Baki Ermis, Pierre Clarot, Pierre-Marc Jodoin, and Venkatesh Saligrama. Activity based matching in distributed camera networks. *Trans. Img. Proc.*, 19(10):2595–2613, October 2010.

[22] A. Gionis, Piotr Indyk, and R. Motwani. Similarity search in high dimensions via hashing. In *Proc. Int. Conf. on Very Large Data Bases*, pages 518–529, 1999.

[23] L. Gorelick, M. Blank, E. Shechtman, M. Irani, and R. Basri. Actions as space-time shapes. *IEEE Trans. Pattern Anal. Machine Intell.*, 29(12):2247–2253, 2007.

[24] T. Hospedales, S. Gong, and T. Xiang. A markov clustering topic model for mining behaviour in video. In *Proc. IEEE Int. Conf. Computer Vision*, pages 1165–1172, 2009.

[25] W. Hu, X. Xiao, Z. Fu, D. Xie, T. Tan, and S. Maybank. A system for learning statistical motion patterns. *IEEE Trans. Pattern Anal. Machine Intell.*, 28(9):1450–1464, 2006.

[26] M. Isard and A. Blake. Condensation – conditional density propagation for visual tracking. *Intern. J. Comput. Vis.*, 29(1):5–28, 1998.

[27] L. Itti and P. Baldi. A principled approach to detecting surprising events in video. In *in Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR*, pages 631–637, 2005.

[28] Y. A. Ivanov and A. F. Bobick. Recognition of visual activities and interactions by stochastic parsing. *IEEE Trans. Pattern Anal. Machine Intell.*, 22(8):852–872, 2000.

[29] F. Jiang, J. Yuan, S. Tsaftaris, and A. Katsaggelos. Anomalous video event detection using spatiotemporal context. *Computer Vision and Image Understanding*, 115(3):323–333, 2011.

[30] P.-M. Jodoin, V. Saligrama, and J. Konrad. Behavior subtraction. In *Proc. SPIE Visual Communications and Image Process.*, volume 6822, pages 10.1–10.12, January 2008.

[31] P. M. Jodoin, V. Saligrama, and J. Konrad. Behavior subtraction. *IEEE Transactions on Image Processing*, 2012.

[32] J. Kim and G. Kristen. Observe locally, infer globally: A space-time MRF for detecting abnormal activities with incremental updates. In *Proc. IEEE Conf. Computer Vision Pattern Recognition*, 2009.

[33] L. Kratz and K. Nishino. Anomaly detection inextremely crowded scenes using spatio-temporal motion pattern models. *CVPR*, 2009.

[34] D. Kuettel, M. Breitenstein, L. Gool, and V. Ferrari. What's going on? discovering spatio-temporal dependencies in dynamic scenes. In *Proc. IEEE Conf. Computer Vision Pattern Recognition*, pages 1951–1958, 2010.

[35] V. Mahadevan, W. Li, V. Bhalodia, and N. Vasconcelos. Anomaly detection in crowded scenes. *CVPR*, 2010.

[36] G. Medioni, I. Cohen, F. Bremond, S. Hongeng, and R. Nevatia. Event detection and analysis from video streams. *IEEE Trans. Pattern Anal. Machine Intell.*, 23(8):873–889, 2001.

[37] R. Mehran, A. Oyama, and M. Shah. Abnormal crowd behavior detection using social force model. In *Proc. IEEE Conf. Computer Vision Pattern Recognition*, 2009.

[38] N. M. Oliver, B. Rosario, and A. P. Pentland. A Bayesian computer vision system for modeling human interactions. *IEEE Trans. Pattern Anal. Machine Intell.*, 22(8):831–843, 2000.

[39] C. Piciarelli, C. Micheloni, and G. Foresti. Trajectory-based anomalous event detection. *IEEE Trans. Circuits Syst. Video Technol.*, 18(11):1544–1554, 2008.

[40] F. Porikli, O. Tuzel, and P. Meer. Covariance tracking using model update based on means on Riemannian manifolds. In *Proc. IEEE Conf. Computer Vision Pattern Recognition*, June 2006.

[41] I. Pruteanu-Malinici and L. Carin. Infinite hidden markov models for unusual-event detection in video. *IEEE Trans. Image Process.*, 17(5):811–821, 2008.

[42] I. Saleemi, K. Shafique, and M. Shah. Probabilistic modeling of scene dynamics for applications in visual surveillance. *IEEE Trans. Pattern Anal. Machine Intell.*, 31(8):1472–1485, 2009.

[43] V. Saligrama and Z. Chen. Video anomaly detection with local statistical aggregates. *CVPR*, 2012.

[44] V. Saligrama and Z. Chen. Video anomaly detection with local statistical aggregates. In *Proc. IEEE Conf. Computer Vision Pattern Recognition*, 2012.

[45] V. Saligrama, J. Konrad, and P.-M. Jodoin. Video anomaly identification: A statistical approach. *IEEE Signal Process. Mag.*, 27:18–33, 2010.

[46] B. Schölkopf, J. Platt, J. Shawe-Taylor, A. Smola, and R. Williamson. Estimating the support of a high-dimensional distribution. *Neural Comput.*, 13(7):1443–1471, 2001.

[47] E. Shechtman and M. Irani. Space-time behavior based correlation or how to tell if two underlying motion fields are similar without computing them? *IEEE Trans. Pattern Anal. Machine Intell.*, 29(11):2045–2056, 2007.

[48] S. Shipman, A. Divakaran, and M. Flynn. Highlight scene detection and video summarization for pvr-enabled high-definition television systems. In *Proc. IEEE Int. Conf. Consumer Electronic*, pages 1–2, 2007.

[49] C. Simon, J. Meessen, and C. DeVleeschouwer. Visual event recognition using decision trees. *Multimedia Tools and Applications*, 2009.

[50] J. Sivic and A. Zisserman. Video Google: A text retrieval approach to object matching in videos. In *Proc. IEEE Int. Conf. Computer Vision*, volume 2, pages 1470–1477, 2003.

[51] X.M. Song and G.L. Fan. Joint key-frame extraction and object segmentation for content-based video analysis. *IEEE Trans. Circuits Syst. Video Technol.*, 16(7):904–914, 2006.

[52] Y-L. Tian, A. Hampapur, L. Brown, R. Feris, M. Lu, A. Senior, C-F. Shu, and Y. Zhai. Event detection, query, and retrieval for video surveillance. In Zongmin Ma, editor, *Artificial Intelligence for Maximizing Content Based Image Retrieval*. Information Science Reference; 1 edition, 2008.

[53] O. Tuzel, F. Porikli, and P. Meer. Region covariance: A fast descriptor for detection and classification. In *Proc. European Conf. Computer Vision*, May 2006.

[54] UCSD. Anomaly detection dataset, 2010. `http://www.svcl.ucsd.edu/projects/anomaly/dataset.htm`.

[55] N. Vaswani, A. Roy-Chowdhury, and R. Chellappa. Shape activity: A continuous state hmm for moving/deforming shapes with application to abnormal activity detection. *IEEE Trans. Image Process.*, 14(10):1603–1616, 2005.

[56] H. Veeraraghavan, N. Papanikolopoulos, and P. Schrater. Learning dynamic event descriptions in image sequences. In *Proc. IEEE Conf. Computer Vision Pattern Recognition*, pages 1–6, 2007.

[57] X. Wang, X. Ma, and E.Grimson. Unsupervised activity perception in crowded and complicated scenes using hierarchical bayesian models. *IEEE Trans. Pattern Anal. Machine Intell.*, 31(3):539–555, 2009.

[58] X. Wang, K. Tieu, and E. Grimson. Learning semantic scene models by trajectory analysis. In *Proc. European Conf. Computer Vision*, pages 111–123, 2006.

[59] T. Xiang and S. Gong. Video behavior profiling for anomaly detection. *IEEE Trans. Pattern Anal. Machine Intell.*, 30(5):893–908, 2008.

[60] C. Yeo, P. Ahammad, K. Ramchandran, and S. Sastr. High speed action recognition and localization in compressed domain videos. *IEEE Trans. Circuits Syst. Video Technol.*, 18(8):1006 – 1015, 2008.

[61] A. Yilmaz and M. Shah. A differential geometric approach to representing the human actions. *Comput. Vis. Image Und.*, 109(3):335–351, 2008.

[62] D. Zhang, D. Gatica-Perez, S. Bengio, and I. McCowan. Semi-supervised adapted hmms for unusual event detection. In *Proc. IEEE Conf. Computer Vision Pattern Recognition*, pages 611–618, 2005.

[63] M. Zhao and V. Saligrama. Anomaly detection with score functions based on nearest neighbor graphs. In *NIPS*, volume 22, 2009.

[64] H. Zhong, J. Shi, and M. Visontai. Detecting unusual activity in video. In *CVPR (2)*, 2004.

[65] X. Zhu, X. Wu, A. Elmagarmid, Z. Feng, and L. Wu. Video data mining: Semantic indexing and event detection from the association perspective. *IEEE Trans. Know. Data Eng.*, 17:665–677, 2005.